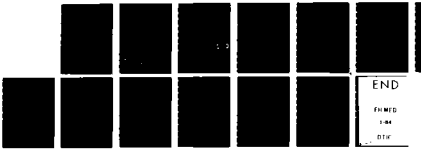
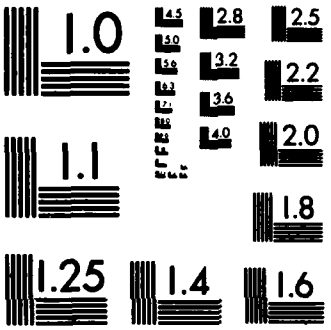
AD-A135 684 COMMENT ON 'GRAPHICAL METHODS FOR ASSESSING LOGISTIC REGRESSION MODELS' B.. (U) CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS S E FIENBERG ET AL. 1/1

UNCLASSIFIED SEP 83 TR-299 N00014-80-C-0637 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

COMMENT ON "GRAPHICAL METHODS FOR ASSESSING LOGISTIC REGRESSION MODELS," BY LANDWEHR PREGIBON, AND SHOEMAKER

by

Stephen E. Fienberg

and

Gail D. Gong*

DEPARTMENT

OF

STATISTICS

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

COMMENT ON "GRAPHICAL METHODS FOR ASSESSING LOGISTIC REGRESSION MODELS," BY LANDWEHR, PREGIBON, AND SHOEMAKER

by

Stephen E. Fienberg

and

Gail D. Gong*

Technical Report No. 299

Department of Statistics

Carnegie-Mellon University

Pittsburgh, PA 15213

September, 1983

---

*Stephen E. Fienberg is Professor of Statistics and Social Sciences and Gail D Gong is Assistant Professor of Statistics at Carnegie-Mellon University, Pittsburgh, PA 15213. The preparation of this paper was partially supported by the Office of Naval Research Contract N00014-80-C-0637 at Carnegie-Mellon University.

The use of graphical methods for diagnostic purposes has an honorable tradition which is rooted in the pioneering work of Anscombe and Tukey and has been developed by Wilk, Gnanadesikan, and a host of others associated with Bell Laboratories. The present paper by Landwehr, Pregibon, and Shoemaker (subsequently referred to as LPS) attempts to modify and extend graphical diagnostic displays that have been developed for ordinary regression to be of use for assessing logistic regression models for binary data. They propose displays for each of the three key components of regression diagnostics: goodness of fit, outlier detection, and model specification. Theirs is a pioneering effort and many useful ideas have emerged from it. We congratulate the authors on the substantial progress they have made to date.

Somewhat fuzzy analogies to linear regression are not sufficient to motivate for us the approaches adopted by LPS. Thus we have attempted to examine critically LPS's diagnostic displays to see if we could determine why in each instance the method works or fails. LPS suggest that the major obstacle in carrying linear regression diagnostics over to the logistic regression setting is the discreteness of binary data. While discreteness may well be a serious problem, we note additional ones. Although our examination is far from complete, we hope it will be a useful supplement to the present paper.

We deal separately with each diagnostic display.

1. LOCAL MEAN DEVIANCE PLOTS

The key idea behind this plot is to focus not on a global measure of goodness-of-fit but rather on local contributions to the fit. The approach is based on an analogy with the linear regression problem with replicated observations where we can partition the sum of squared residuals (SS) into a pure-error SS and a lack-of-fit SS. LPS claim that for the logistic regression setting that, "if there are exact replicates in the data, the pure-error component of the deviance is easily obtained." This is true in a sense, but it is somewhat deceptive.

As in the case of the linear regression problem, when we have exact replicates in an observational study the definition of pure error depends on two assumptions: (i) independence

of observations, and (ii) a correct choice of the explanatory variables for inclusion in the model (but not necessarily the functional form). Independence is a major problem in many applications although it does seem reasonable in LPS's major example. The specification of relevant explanatory variables is more serious but we have no advice for dealing with it except to proceed conditionally, which is in effect what LPS do with their example. Thus despite our note of caution, we find the idea of using a pure error component appealing.

Now, LPS are dealing with the case where there are not necessarily exact replicates, and they do not really attempt to partition the deviance. What then are they doing? Through the use of their clustering algorithm, in effect they are partitioning the factor space (i.e. the m dimensional space of explanatory variables) into K distinct regions. Then they attempt to check whether the logistic regression in each region is the same as the global logistic regression by fitting the model

$$\text{logit } (p) = Z\gamma + X\beta, \tag{1.1}$$

where $Z_{ik} = 1$ or 0 according to whether the ith observation is or is not in the kth region. We wish to determine if the model is the same in all regions, i.e. $\gamma = 0$ or

$$\text{logit } (p) = X\beta, \tag{1.2}$$

and we can test for this directly using a conditional likelihood ratio test of model (1.2) versus model (1.1). If N is large relative to K such a test, and if the regions were preformed, the conditions of Haberman (1974) would seem to be satisfied and a χ^2 reference distribution with K-1 degrees of freedom would be appropriate. (Tsiatis, 1980, proposes a similar test but uses a Wald-like quadratic form statistic.) LPS eschew a formal test, and approach the issue graphically by examining the contributions to such a conditional test statistic by region of the factor space, in a cumulative form. This seems reasonable if we don't have preformed groups of data points, and if we have relied on a clustering algorithm of the sort suggested by LPS.

How well does the LPS graphical approach work? We can explore the answer to this question best in the case where each cluster or region of the factor space consists of exact replicates. As Jennings (1982) notes, under the null hypothesis of model (1.2), the expectation

of the local mean square deviance for the kth group, $\bar{D}_k/(N_k-1)$, should be approximately the same as the expectation of the global mean deviance, $\bar{D}$. Jennings has carried out calculations for some simple examples that suggest that the local mean deviance overestimates the global mean deviance for p near 0.5 and underestimates it for p near 0 or 1. Thus the order in which the groups are added to LPS's running estimate of local mean deviance may have a substantial impact on what we see in the graphical display. As a consequence we are unable to share LPS's enthusiasm for this display.

A simple alternative to the LPS approach, and one we have found successful in practice, is suggested by the preceding discussion. Convert each of the m explanatory variables into sets of categories, and restructure the data in the form of an (m+1)-dimensional contingency table with an m-dimensional fixed margin (Bishop, Fienberg and Holland, 1975; Fienberg, 1980). Now examine the fit of the logistic regression model using, say, the means of the explanatory variables in each cell. To examine local variation we can use the generalized residual approach of Haberman (1976). This contingency table structuring can also be used to explore directly nonlinear effects of the explanatory variables and interactions.

2. EMPIRICAL PROBABILITY PLOTS

Both this plot and the next one involve the adaptation of the linear regression notion of standardized residuals. Here, LPS begin by arguing for the use of the deviance contribution, $d_i = d(\hat{p}_i; y_i)$, standardized by its approximate standard error. We see little justification considering the use of a χ^2_1 reference distribution in this setting given that, for extreme values of p_i, d_i^2 tends to a 2-point distribution and Haberman's (1974) conditions are not met (e.g. see the discussion in Jennings, 1982).

LPS's alternative is to use the residuals $y_i - \hat{p}_i$ and a simulation procedure. The procedure here is evocative of one proposed by Atkinson (1981, 1982), in which he presents half normal plots of jackknife residuals and a modified version of Cook's distance statistic, using 19 samples simulated with random normal y's and a matrix of explanatory variables the same as

that of the data. Because both statistics that Atkinson uses are functions of least squares residuals divided by an estimate of scale, the values of the parameters of the linear model in the simulation *do not* matter. This is not the case in LPS's simulation for their empirical probability plot. Even if they had standardized their residuals the values of the parameters in the logistic regression model, β, *do* matter. LPS attempt to finesse this problem by using $\hat{p}_i$ in place of $\hat{p}_i$ for the simulated Bernoulli variates, but this simply disguises the dependence of the simulation on the p_i. Thus we are skeptical about attaching the usual interpretation to the confidence coefficients used throughout Section 4 of LPS.

In the concluding discussion, LPS mention that the simulations for the empirical probability plots are in the same spirit as the bootstrap. In fact the simulations used to get the distribution of the ordered residuals are a form of bootstrap. We illustrate the bootstrap argument using LPS's Example 3. We assume that (X_1,Y_1), ..., (X_{100},Y_{100}) are independent and identically distributed (iid) from some distribution F, where F specifies that $Y|X$ is Bernoulli with probability of success $P(X)$ such that logit $P(X) = X\beta$. Then we form the residuals $Y - \hat{Y}$ where logit $\hat{Y} = X\hat{\beta}$. If we knew F, we could in principle simulate to get the distribution of the ordered residuals under the assumed model F. Not knowing F, we substitute an estimate. The appropriate estimate here is $\hat{F}$ which specifies that the probability of success is $\hat{Y}$. We comment further on the use of the bootstrap in the next section.

3. PARTIAL RESIDUAL PLOTS

In the case of ordinary linear regression, collinearity in the design space can hide a sought-after relationship from partial residuals. More specifically, if the underlying model is

$$E(Y) = X\beta + g(z), \tag{3.1}$$

then partial residuals fail to detect the part of $g(z)$ which is in the column space of X. We expect similar problems to hold in the logistic regression case.

LPS motivate their definition of partial residuals by considering an analogy to ordinary linear regression. Here, we first give a simple calculation to show why partial residuals sometimes

work, and then we give an example to show when partial residuals can fail.

Suppose Y is Bernoulli with probability of success P(z) such that

$$\text{logit } P(z) = \alpha + g(z), \tag{3.2}$$

where z is real. We fit the logistic model

$$\text{logit } P_1(z) = \alpha + \gamma z. \tag{3.3}$$

Linear logistic regression estimates P_1 in (3.3) when what we are after is P in (3.2). Suppose we define the residual logit

$$r_{logit} = \text{logit } P - \text{logit } P_1 = \log\left(\frac{P}{1-P} \cdot \frac{1-P_1}{P_1} \right). \tag{3.4}$$

Using the first order Taylor approximation for the logarithm, we have

$$\begin{aligned} r_{logit} &= \text{logit } P - \text{logit } P_1 \\ &\approx \frac{P}{1-P} \cdot \frac{1-P_1}{P_1} - 1 \\ &= \frac{P - P_1}{P_1 (1-P)} \\ &\approx \frac{P - P_1}{P_1(1-P_1)}. \end{aligned} \tag{3.5}$$

The approximations hold if the ratios

$$\frac{P}{P_1} \approx 1, \quad \frac{1-P}{1-P_1} \approx 1. \tag{3.6}$$

Solving for g(z) in (3.2) and (3.3), we get

$$g(z) = r_{logit} + \gamma z = \frac{P-P_1}{P_1(1-P_1)} + \gamma z. \tag{3.7}$$

If we define

$$G(z) = \frac{Y-P_1}{P_1(1-P_1)} + \gamma z, \tag{3.8}$$

then

$$E(G(z)|z) \approx g(z). \tag{3.9}$$

Substituting estimates for P_1 and γ in (3.8) gives r_{par} as defined by LPS. This simple calculation shows that when (3.6) holds, the conditional expectation of r_{par} given z is approximately $g(z)$. When z is discrete and the number of observations at each possible value of z is large, we can average the partial residuals at each value of z to get an estimate of $g(z)$. When z is continuous, we cannot take averages at each possible value of z, but we can smooth.

We now give a counterexample in which LPS's partial residual display fails. Suppose logit $P(z) = \log(1+z)$ where the possible values of z are

$$0,\ 2,\ 4,\ 6,\ 8,\ 10,\ 20,\ 30,\ 40,\ 50.$$

It turns out that

$$\text{logit } P_1 = \alpha + \gamma\ z,$$

where $\alpha = 0.85$ and $\gamma = 0.10$. Figure 1 shows a scatterplot of $G(z)$. For each z, $G(z)$ can take two possible values depending on whether $Y=1$ or $Y=0$. At these two values, the number of dashes reflect the true proportions of successes $P(z)$ and failures $1-P(z)$. We may think of Figure 1 as the partial residual plot gotten by looking at an infinite number of observations at each z. Comparing $E(G(z)|z)$ and $g(z) = \log(1+z) - \alpha$, we see that partial residuals fail when z exceeds 30. In Figure 2, we calculate

$$\frac{P}{P_1} \text{ and } \frac{1-P_1}{1-P},$$

for $z = 30,\ 40,\ 50$, and we see that for these values of z, the approximation (3.6) simply does not hold.

Our final comment on partial residuals is one on assessment. In the breast cancer example, LPS examine numerous partial residual plots to obtain a rather complicated model involving 7 parameters. Assessing the fit of this model is a difficult but important problem. There are

two questions we might ask. If new patients were observed, would the model afford a good rule for predicting their outcome? If the experiment were performed again with a new sample of 306 patients, would LPS conclude the same model? If we could automate the thought processes that produce say, the conclusion that LPS's Figure 8 shows a cubic dependence, we could use the bootstrap to help answer these questions. (See Gong 1982a, 1982b for the use of the bootstrap in assessing another complicated prediction rule.)

We include here just an idea of how the bootstrap can help. Look at LPS's Figure 8. If we repeated the experiment, observing a new sample of 306 patients, would we get a picture similar to their Figure 8, and would LPS arrive at the same conclusion of a cubic dependence? Suppose the data (X_1,Y_1), ..., (X_{306},Y_{306}) were iid with distribution F. If we knew F, we could in principle generate a new sample of patients. Since we don't know F, substitute an estimate, the empirical distribution $\hat{F}$ which puts mass 1/306 at each observation. The resulting sample is a bootstrap sample. Our Figure 3 shows the partial residual smooths of 10 bootstrap samples. (Performing 25 bootstraps gives similar results but a more confusing picture.) How do we interpret this picture? The smooth based on the original sample is our estimator of the nonlinear relation of logit P on age. The smooths based on the bootstrap samples tell us about the variability of that estimator. Since the shape of the bootstrap smooths all tend to be similar, we have some confidence in the original sample smooth as an estimator for the shape of the nonlinear relation of logit P on age.

Figure 4 shows the partial residual smooths of 10 bootstraps for the second covariate, the year of surgery. The shape of these smooths are also very similar, indicating that the partial residual smooth of the original sample, also given in Figure 4 is a good estimator for the nonlinear dependence of logit P on year of surgery. There is an interesting decrease in survival for surgery during 1965.

Figure 5 shows the smooths of 10 bootstraps for the third covariate, the number of nodes. The shape of the smooths for small number of nodes remains constant throughout the

bootstraps, while the shape for large number of nodes is highly variable. If logit P does depend on number of nodes through $-\log(1+z)$, we have a situation similar to our counterexample. For z large, the difference between G(z,Y=1) and G(z,Y=0) is astronomical, so that observing a few more Y=1's can pull the smooth way up, while observing a few more Y=0's can pull the smooth down. The high variability in shape is ultimately tied to the fact that there are just not very many patients with large number of nodes.

4. ADDITIONAL COMMENTS

Although the amount of calculation required to produce the graphical displays proposed by LPS seems at first blush immense, the plots can in fact be generated with relative ease using a statistical software package such as Bell Laboratories' "S". This is a major virtue, and argues for further efforts to build up on LPS's pioneering work. Variants on and alternatives to their three graphical displays can be explored with the same ease as can the original displays.

In future work in this area we would argue for somewhat less reliance on fuzzy analogies to linear regression techniques, and we would focus more *directly* on the theory that underlies the problem at hand. Of potential importance for the logistic regression setting is Jennings' (1982) measure of inference adequacy, a measure closely related to the curvature measures of Bates and Watts (1980). His approach has implications both for model specification and for the form of parametrization of the logistic regression model to achieve computational efficiency and inferential stability.

The use of kernel-based density estimation in the development of grapical displays for binary response models also bears scrutiny (e.g. see Titterington, 1980, and Titterington et al., 1981). Although such an approach has been advocated as "non-parametric" (e.g. see Copas, 1983) we see its major strengths as (a) the smoothness of the resulting probabilities, and (b) the checks it might provide for the examination of model adequacy and local variation.

Developing *useful* graphical displays is, in general, a difficult task. LPS have demonstrated how ideas for diagnostic displays from ordinary linear regression can be adapted to the binary

response setting. In our examination of their work we have pointed out both strengths and weaknesses in their methodology. LPS have opened the gates on a research field that has yet to fully bloom. Much work still needs to be done before we can reap a harvest of useful methodology.

REFERENCES

Atkinson, A.C. (1981), "Two graphical displays for outlying and influential observations," *Biometrika*, **68**, 13–20.

Atkinson, A.C. (1982), "Regression diagnostics, transformations and constructed variables (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, **44**, 1–36.

Bates, D.M. and Watts, D.G. (1980), "Relative curvature measures of nonlinearity (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, **40**, 1–25.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Mass.: MIT Press.

Copas, J.B. (1983), "Plotting p against x," *Applied Statistics*, **32**, 25–31.

Fienberg, S.E. (1980), *The Analysis of Cross-classified Categorical Data* (2nd ed.), Cambridge, Mass.: MIT Press.

Gong, G.D. (1982a), "Some ideas on using the bootstrap in assessing model variability," in *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, eds. K.W. Heiner, R.S. Sacher, J.W. Wilkinson, New York: Springer-Verlag, 169–173.

Gong, G.D. (1982b), Cross-validation, the jackknife, and the bootstrap: excess error estimation in stepwise logistic regression, Ph.D. Thesis, Department of Statistics, Stanford University.

Haberman, S.J. (1974), *The Analysis of Frequency Data*, Chicago, Ill.: Univ. of Chicago Press.

Haberman, S.J. (1976), Generalized residuals for loglinear models, *Proceedings of the Ninth International Biometrics Conference, Boston*, 104–122.

Jennings, D.E. (1982), Inference and diagnostics for logistic regression, Ph.D. Thesis, School of Statistics, University of Minnesota.

Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984), "Graphical methods for assessing logistic regression models," *Journal of the American Statistical Association*, **79**, (forthcoming).

Titterington, D.M. (1980), "A comparative study of kernel-based density estimates for categorical data," *Technometrics*, **22**, 259–268.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., and Gelpke, G.J. (1981), "Comparison of discrimination techniques applied to a complex data set of head injury patients (with discussion)," *Journal of the Royal Statistical Society*, Ser. A, 145–174.

Tsiatis, A.A. (1980), "A note on a goodness-of-fit test for the logistic regression model," *Biometrika*, **67**, 250–251.

Figure 1. The Partial Residual Plot for the Counterexample.

The dashes form a scatterplot of G(z). The curve through this scatterplot is E(G(z)|z) which estimates the true function g(z) indicated by the triangles.

Figure 2. Understanding the Failure of Partial Residuals

Condition (3.6) fails to hold for z = 30, 40, 50 in the counterexample.

z	P	P_1	P/P_1	$(1-P_1)/(1-P)$
30	.969	.980	.99	.65
40	.976	.993	.98	.29
50	.981	.997	.98	.16

Figure 3. Bootstraps for Age

The partial residual smooths of 10 bootstrap samples together with that of the original sample of the covariate x_1 = age.

Figure 4. Bootstraps for Year

The partial residual smooths of 10 bootstrap samples together with that of the original sample of the covariate x_2 = year of surgery.

Figure 5. Bootstraps for Nodes

The partial residual smooths of 10 bootstrap samples together with that of the original sample of the covariate x_3 = number of nodes.

FIGURE 1

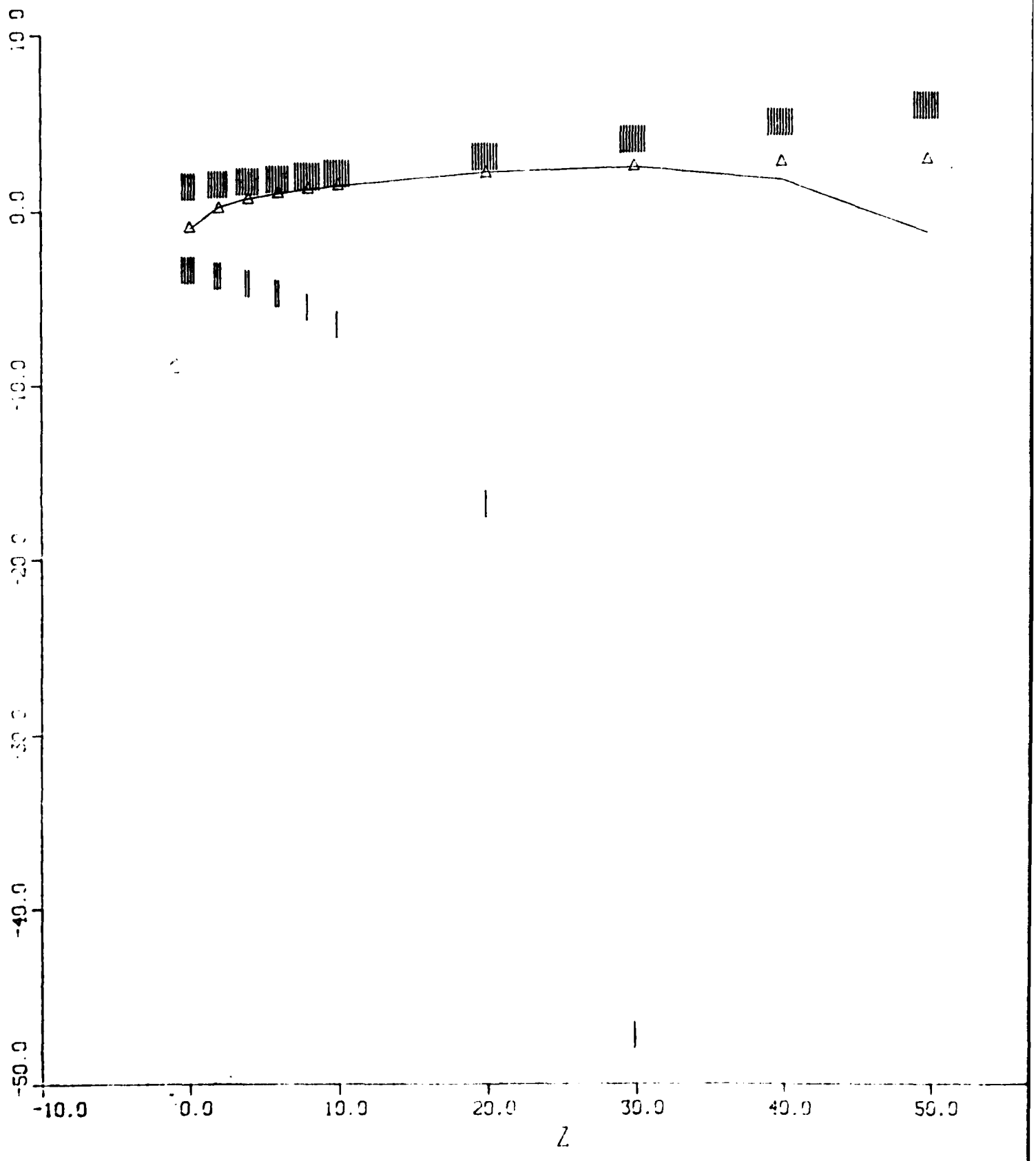

PARTIAL RESIDUAL PLOT FOR COUNTEREXAMPLE
10.0
0.0
-10.0
-20.0
-30.0
-40.0
-50.0
-10.0
0.0
10.0
20.0
30.0
40.0
50.0
Z

FIGURE 3

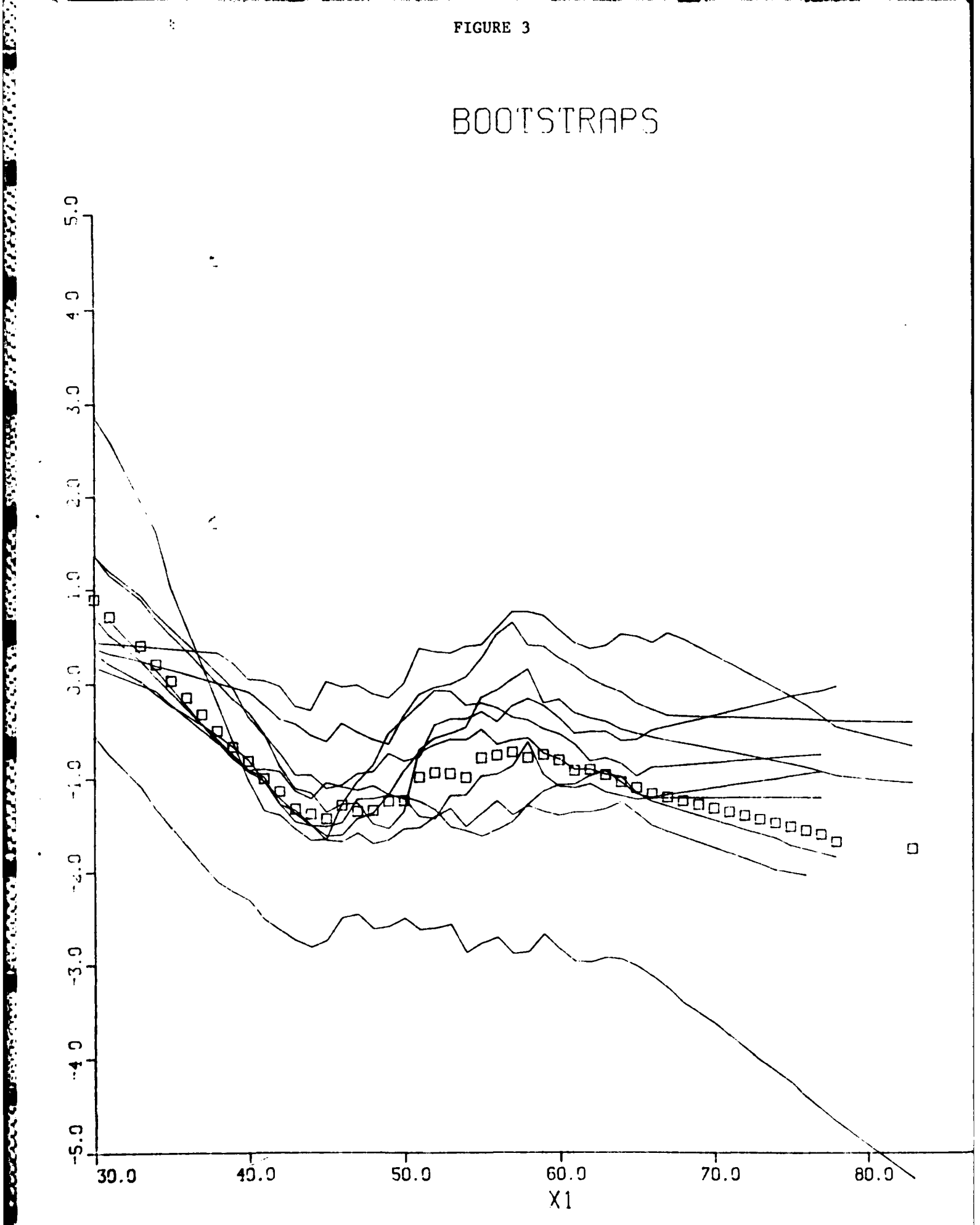

BOOTSTRAPS
5.0
4.0
3.0
2.0
1.0
-1.0
-2.0
-3.0
-4.0
-5.0
30.0
40.0
50.0
60.0
70.0
80.0
X1

FIGURE 4

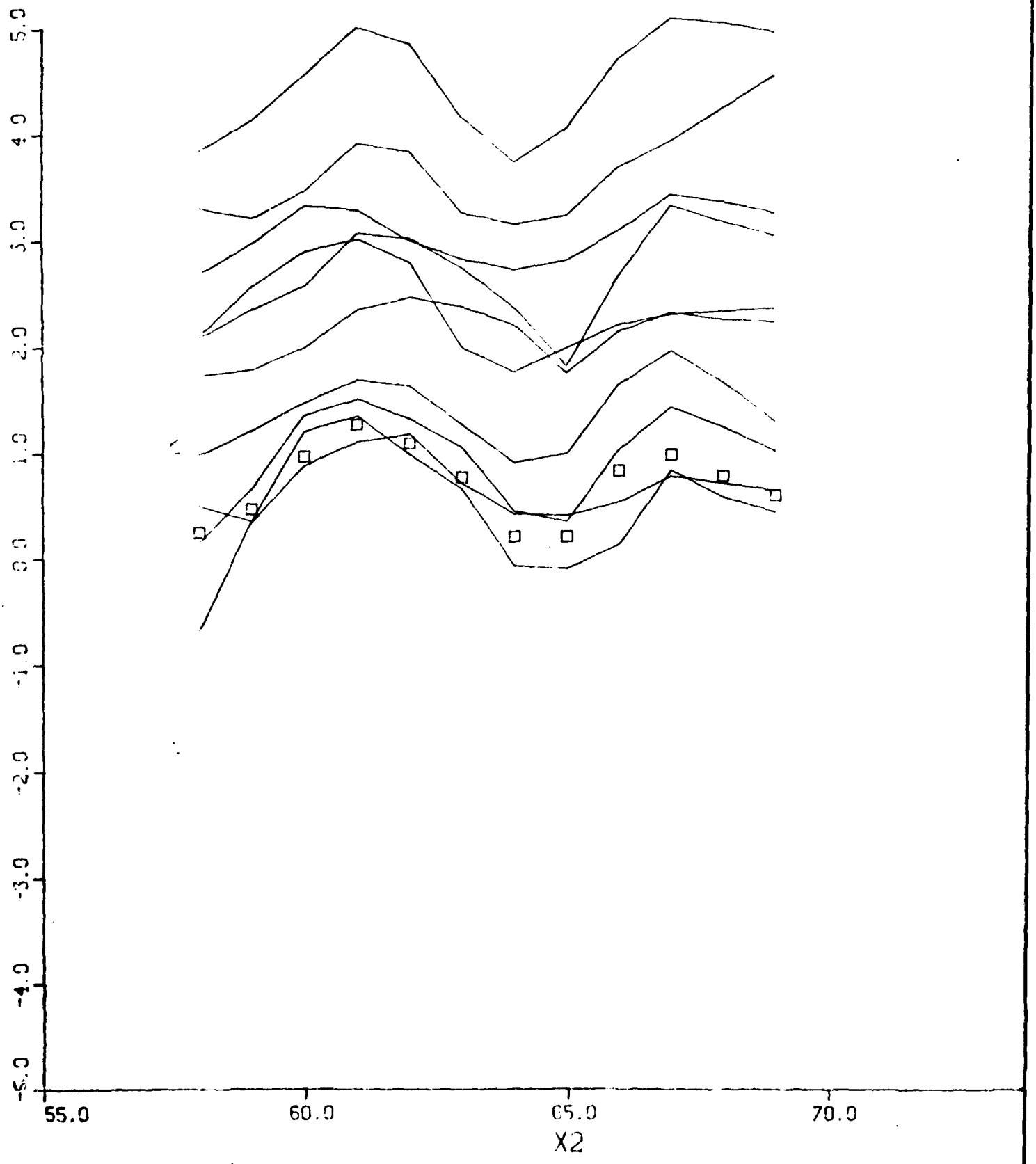
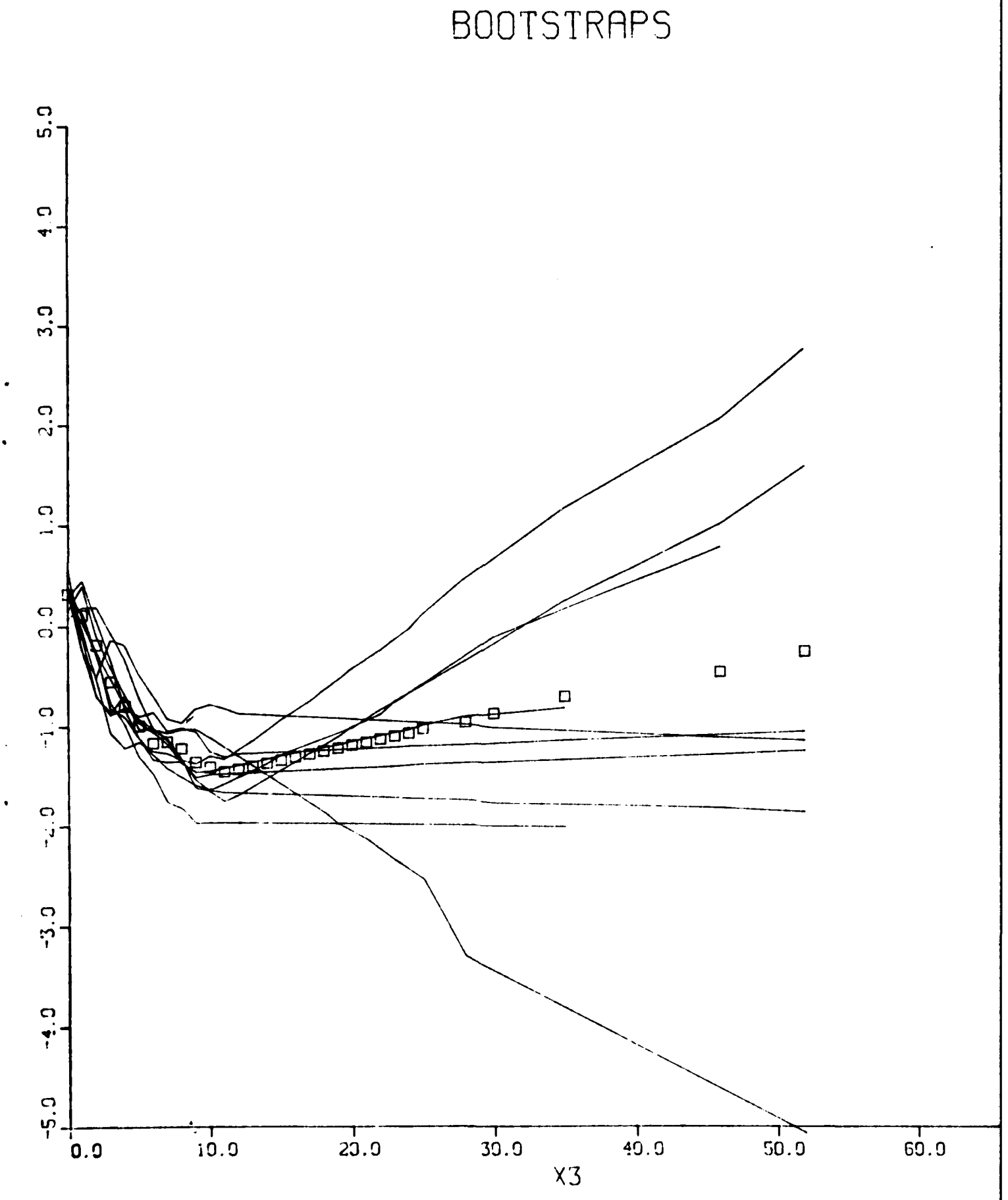


FIGURE 5


BOOTSTRAPS
5.0
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
0.0
10.0
20.0
30.0
40.0
50.0
60.0
X3

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 299	2. GOVT ACCESSION NO. AD-A135684	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Comment on "Graphical Methods for Assessing Logistic Regression Models," by Landwehr, Pregibon, and Shoemaker		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Stephen E. Fienberg Gail D. Gong		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE September, 1983 13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

1-84

DTIC